IBM Developer
SKILLS NETWORK

# Winning Space Race with Data Science

Suhas SD
08 sept 2024

# Outline

- Executive Summary
- Introduction
- Methodology
- Results
- Conclusion
- Appendix

# Executive Summary

- This report presents a detailed analysis of SpaceX’s launch site locations and Falcon 9 landing predictions, leveraging geospatial insights, predictive modeling, and interactive dashboards.
- **Methodologies:**
- **Geospatial Analysis:**
  - **Goal:** Assess proximity of launch sites to railways, highways, coastlines, and cities.
  - **Tools:** Used Folium for mapping and distance calculation, revealing close proximities to key infrastructure.

# Executive Summary

- **Predictive Modeling:**
  - **Goal:** Predict Falcon 9 landing success.
  - **Models Tested:** Logistic Regression, SVM, Decision Tree, K-Nearest Neighbors.
  - **Best Model:** SVM with 84.8% cross-validation accuracy and 83.3% test accuracy, optimized using GridSearchCV.
- **Interactive Dashboard:**
  - **Goal:** Visualize SpaceX launch data dynamically.
  - **Features:** Created with Plotly Dash, including filters, pie charts, and scatter plots for exploring success rates and payload correlations.

# Executive Summary

**Results:**

- **Geospatial Analysis:** Confirmed strategic placement of launch sites near critical infrastructure.
- **Predictive Modeling:** SVM emerged as the most accurate model for landing success predictions.
- **Dashboard:** Provided effective visualization and interaction with launch data, enhancing analysis capabilities.
- This concise overview integrates key findings and methodologies, offering a snapshot of the analysis and its implications.

# Introduction

- **Project Background and Context:** This project examines SpaceX launch sites and Falcon 9 landing predictions to optimize site selection and mission planning through spatial and predictive analysis.
- **Problems to Address:**
- Assess the proximity of launch sites to railways, highways, coastlines, and cities.
- Identify the best model for predicting Falcon 9 landing success.
- Create a dashboard for visualizing launch success and payload impacts.
- The goal is to leverage data for improved launch site utilization and landing accuracy.

Section 1

# Methodology

# Methodology

## Executive Summary

- Data collection methodology:
  - `Gathered data from SpaceX launch records and related sources.`
- Perform data wrangling
  - `Cleaned, wrangled, and transformed data for analysis and visualization.`Performed exploratory data analysis (EDA) using visualization and SQL
- Perform interactive visual analytics using Folium and Plotly Dash
- Perform predictive analysis using classification models
  - `Develope classification models, optimized hyperparameters, and assessed performance using accuracy metrics.`

# Data Collection

- **Source Identification**
  - **Data Sources**: Identified relevant datasets from online repositories and internal sources.
- **Data Acquisition**
  - **Download**: Retrieved data files from specified URLs or databases.
  - **APIs**: Used APIs for real-time data extraction when needed.
- **Data Storage**
  - **Local Storage**: Saved datasets to local directories or cloud storage.
  - **Database**: Imported data into relational databases for structured access.

```
Start
  |
Identify Data Sources
  |
  V
Retrieve Data (Download, APIs)
  |
  V
Store Data (Local, Database)
  |
  V
End
```

# Data Collection – SpaceX API

**1.API Endpoint Identification**
- •**Endpoints**: Selected SpaceX REST API endpoints for launch data.
- •**Documentation**: Referenced SpaceX API documentation for available endpoints.

**2.REST API Calls**
- •**GET Requests**: Used GET requests to retrieve launch records and rocket details.
- •**Authentication**: Managed any required API keys or tokens.

**3.Data Retrieval**
- •**Format**: Collected data in JSON format.
- •**Integration**: Integrated data into the analysis environment.

**4.Data Storage**
- •**Local Files**: Saved retrieved data to local CSV or JSON files.
- •**Database**: Optionally imported data into a relational database for further analysis.

Start
|
Identify API Endpoints
V
Make REST API Calls (GET Requests)
V
Retrieve Data (JSON Format)
V
Store Data (Local Files, Database)
V
End

# Data Collection - Scraping

•**Identify Target Sites**
Selected relevant websites and pages.
•**Scraping Tools**
Used BeautifulSoup and Scrapy for data extraction.
•**Data Extraction**
Parsed HTML to extract needed data.
•**Store Data**
Saved data in CSV/JSON formats.

Start
|
Identify Sites
|
Use Scraping Tools
|
Fetch & Parse HTML
|
Extract & Store Data
|
End

# Data Wrangling

- **Data Processing Steps**
- **Cleaned Data**: Addressed missing values and duplicates.
- **Transformed Data**: Standardized formats and normalized values.
- **Integrated Data**: Merged datasets and aligned schemas.
- **Engineered Features**: Created new features and encoded categorical variables.

Start
|
Data Cleaning
|
Data Transformation
|
Data Integration
|
Feature Engineering
|
End

# EDA with Data Visualization

- **Charts and Their Purposes**
- **Pie Charts**: Visualized categorical success rates by launch site.
- **Scatter Plots**: Analyzed payload mass vs. launch success.
- **Histograms**: Examined distribution of payload masses and other variables

# EDA with SQL

- **Charts and Their Purposes**
- **Pie Charts**: Visualized categorical success rates by launch site.
- **Scatter Plots**: Analyzed payload mass vs. launch success.
- **Histograms**: Examined distribution of payload masses and other variables

# Build an Interactive Map with Folium

- **Map Objects Created and Their Purpose**
- **Markers**: Indicated proximity to cities, railways, and highways.
- **Polylines**: Visualized distances between launch sites and key locations.
- **Circles**: (if used) Represented proximity zones around launch sites.
- **Purpose**
  These objects provided insights into geographic relationships and proximities.

# Build a Dashboard with Plotly Dash

- **Plots/Graphs Added**
- **Pie Chart**: Displayed success/failure rates by launch site.
- **Scatter Plot**: Showed correlation between payload mass and launch success.
- **Interactions**
- **Dropdown Menu**: Allowed selection of specific launch sites.
- **Range Slider**: Enabled filtering payload mass for detailed analysis.
- **Purpose**
  These visualizations and interactions provided a comprehensive view of launch success and payload impacts.

# Predictive Analysis (Classification)

- **Model Development Process**
- **Built Models**: Logistic Regression, SVM, Decision Tree, KNN.
- **Evaluation**: Used cross-validation and test accuracy.
- **Improvement**: Tuned hyperparameters using GridSearchCV.
- **Best Model**: SVM with kernel='sigmoid' and C=1 achieved highest accuracy.

# Results

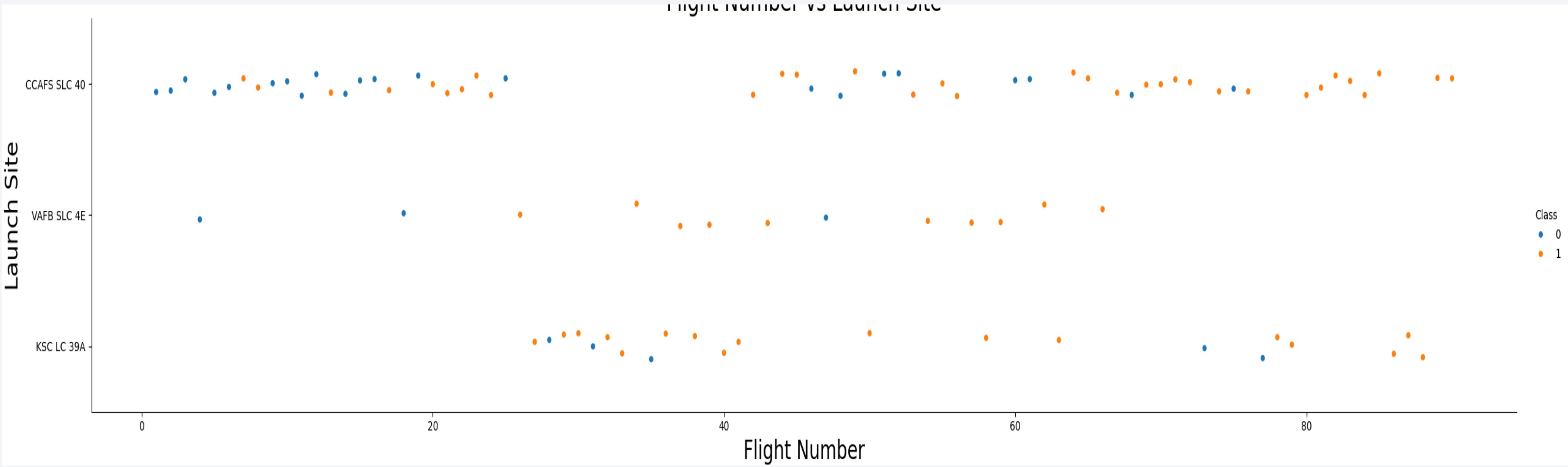
- **Exploratory Data Analysis (EDA) Results**
- **Key Findings:** Identified that launch sites are close to railways, highways, and coastlines, with varying distances from cities. Interactive charts showed success rates and payload correlations.
- **Interactive Analytics Demo**
- **Screenshots:** Demonstrated interactive maps with launch site proximities to cities, railways, highways, and coastlines, and dashboards illustrating launch success by site and payload vs. success correlations.
- **Predictive Analysis Results**
- **Model Performance:** Best model, SVM with kernel='sigmoid', achieved an accuracy of 83.33% on test data.

Section 2

# Insights drawn from EDA

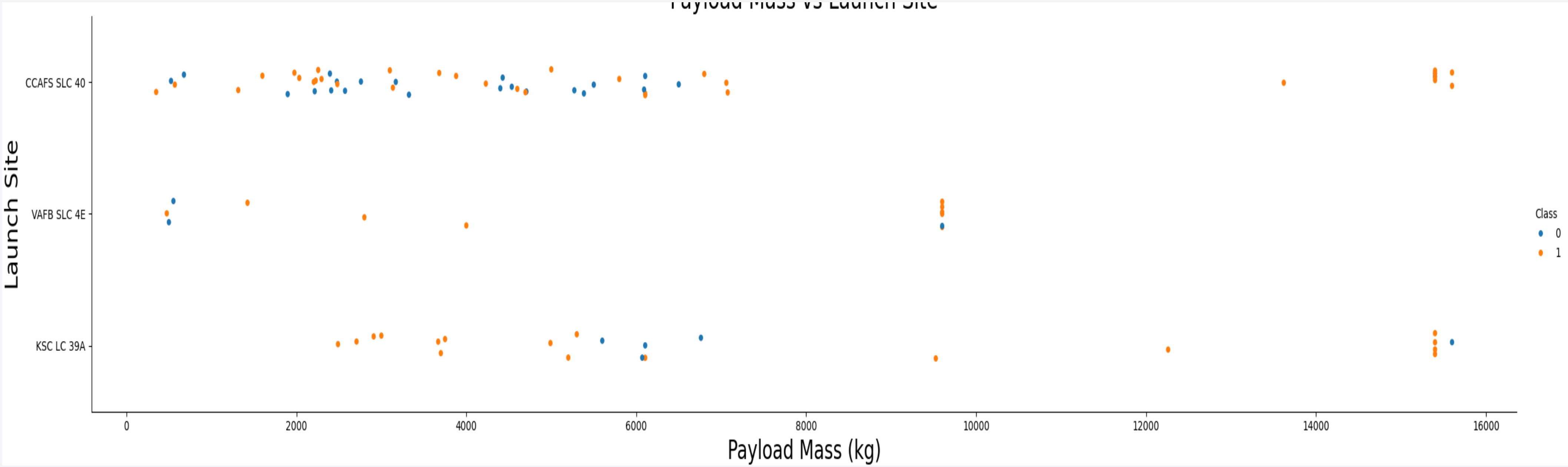
# Flight Number vs. Launch Site

- Scatter plot showing flight numbers against launch sites, indicating the number of launches per site.



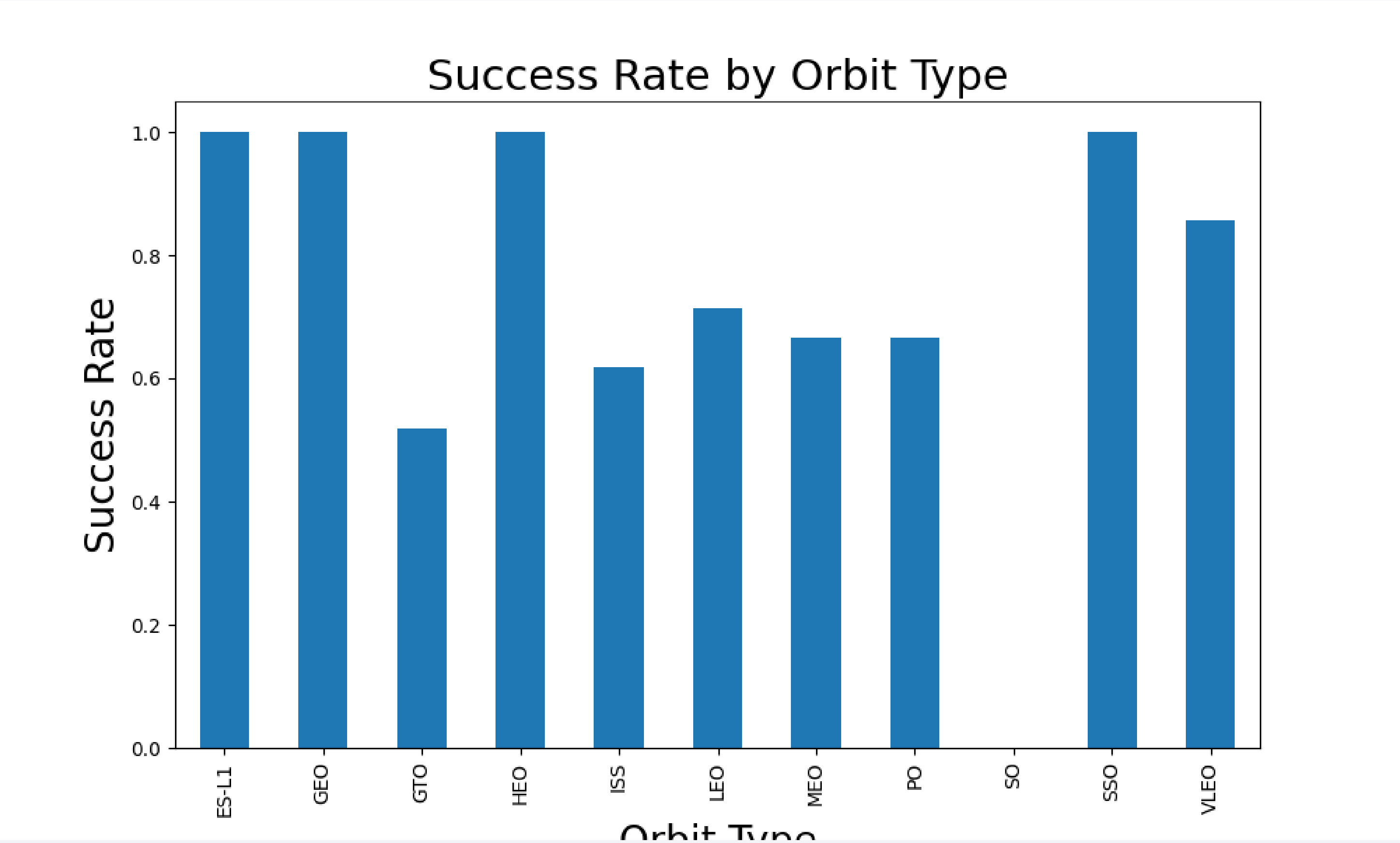
# Payload vs. Launch Site

- Bar chart showing the correlation between payload and launch site, analyzing payload distribution.



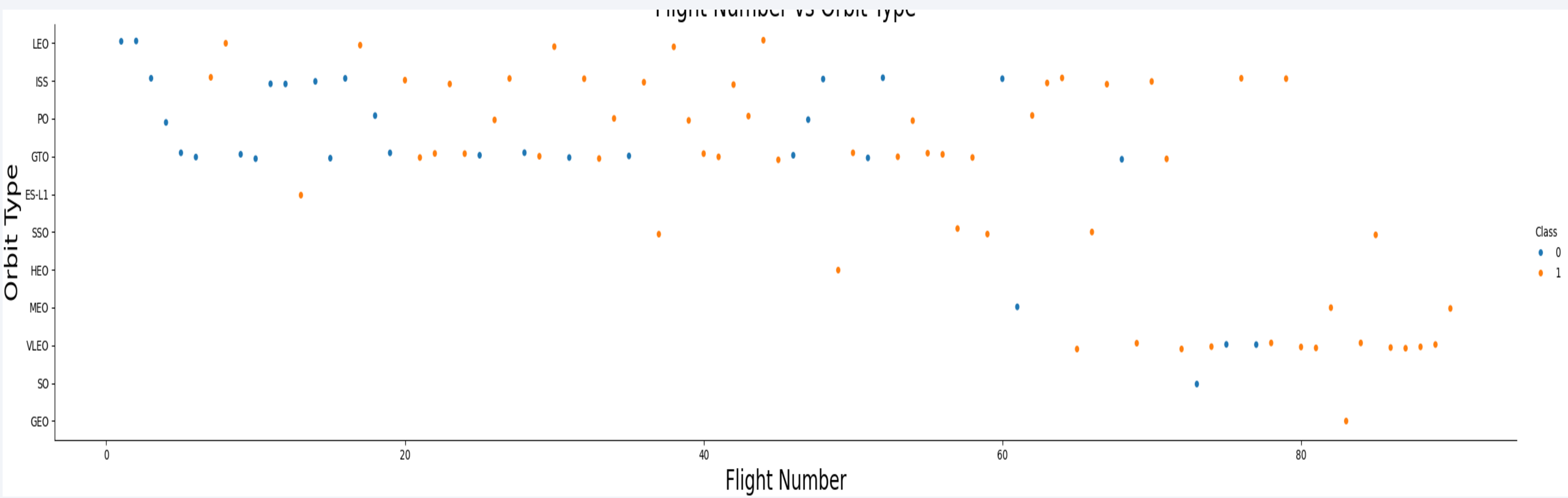
# Success Rate vs. Orbit Type

- Bar chart representing success rates across different orbit types.



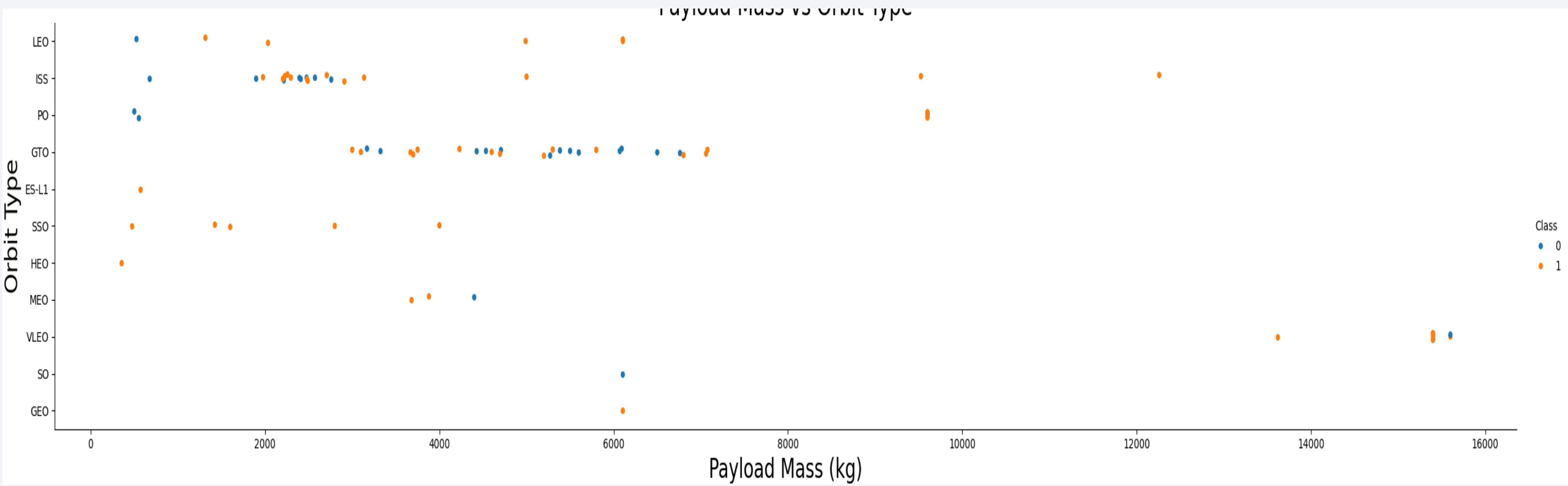
# Flight Number vs. Orbit Type

- Scatter plot of flight number versus orbit type to study launch patterns by orbit.



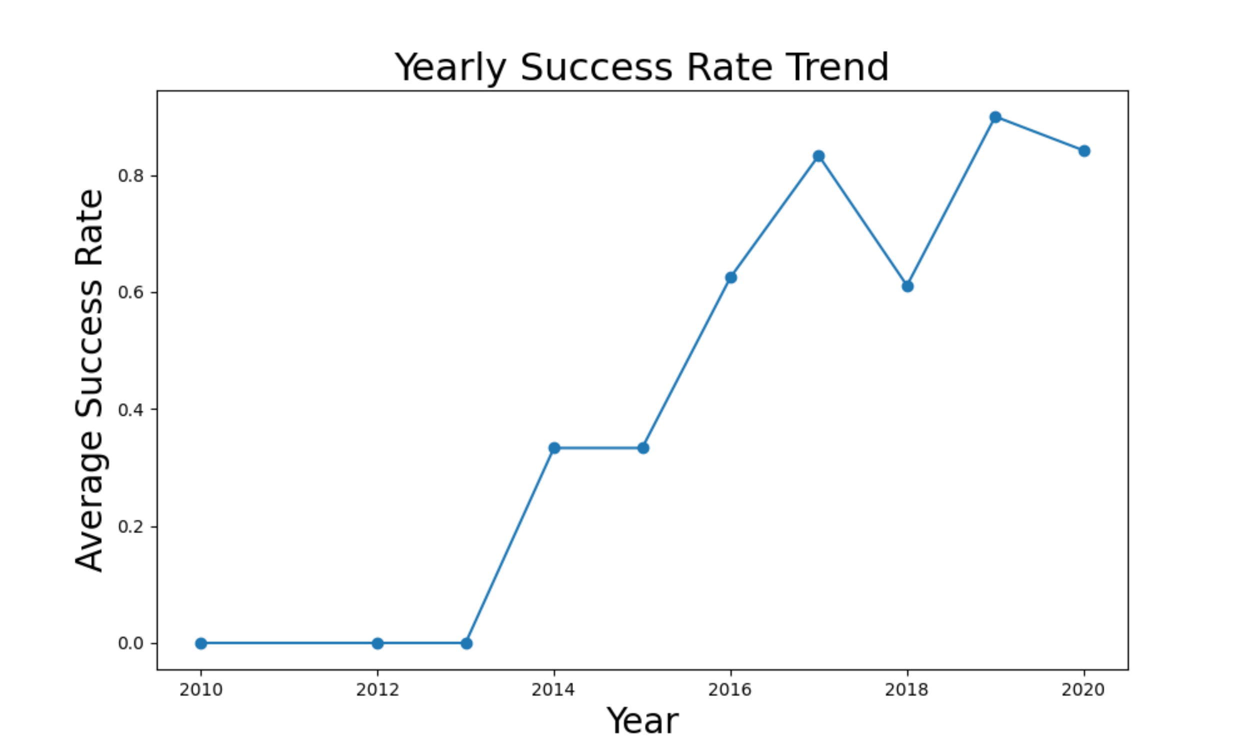
# Payload vs. Orbit Type

- Scatter plot showing payload against orbit type, helping visualize payload capacity trends.



# Launch Success Yearly Trend

- Line chart showing the trend of yearly launch successes, highlighting SpaceX's success over time.



# All Launch Site Names

- %%sqlSELECT DISTINCT "Launch_Site"FROM SPACEXTABLE;
- SQL query to fetch all distinct launch site names, listing available launch sites.

| Launch_Site |
|---|
| CCAFS LC-40 |
| VAFB SLC-4E |
| KSC LC-39A |
| CCAFS SLC-40 |

# Launch Site Names Begin with 'CCA'

- %%sqlSELECT *FROM SPACEXTABLEWHERE "Launch_Site" LIKE 'CCA%‘

| Date | Time (UTC) | Booster_Version | Launch_Site | Payload | PAYLOAD_MASS__KG_ | Orbit | Customer | Mission_Outcome | Landing_Outcome |
|---|---|---|---|---|---|---|---|---|---|
| 2010-06-04 | 18:45:00 | F9 v1.0 B0003 | CCAFS LC-40 | Dragon Spacecraft Qualification Unit | 0 | LEO | SpaceX | Success | Failure (parachute) |
| 2010-12-08 | 15:43:00 | F9 v1.0 B0004 | CCAFS LC-40 | Dragon demo flight C1, two CubeSats, barrel of Brouere cheese | 0 | LEO (ISS) | NASA (COTS) NRO | Success | Failure (parachute) |
| 2012-05-22 | 7:44:00 | F9 v1.0 B0005 | CCAFS LC-40 | Dragon demo flight C2 | 525 | LEO (ISS) | NASA (COTS) | Success | No attempt |
| 2012-10-08 | 0:35:00 | F9 v1.0 B0006 | CCAFS LC-40 | SpaceX CRS-1 | 500 | LEO (ISS) | NASA (CRS) | Success | No attempt |
| 2013-03-01 | 15:10:00 | F9 v1.0 B0007 | CCAFS LC-40 | SpaceX CRS-2 | 677 | LEO (ISS) | NASA (CRS) | Success | No attempt |

# Total Payload Mass

- %%sqlSELECT SUM("PAYLOAD_MASS__KG_") AS Total_Payload_MassFROM SPACEXTABLEWHERE "Customer" = 'NASA (CRS)';
- **Total_Payload_Mass =** 45596
- SQL query to calculate the total payload mass delivered by NASA missions.

# Average Payload Mass by F9 v1.1

- %%sqlSELECT AVG("PAYLOAD_MASS__KG_") AS Avg_Payload_MassFROM SPACEXTABLEWHERE "Booster_Version" = 'F9 v1.1';
- **Avg_Payload_Mass =** 2928.4
- SQL query to calculate the average payload mass for F9 v1.1 missions.

# First Successful Ground Landing Date

- %%sqlSELECT MIN("Date") AS First_Successful_LandingFROM SPACEXTABLEWHERE "Landing_Outcome" = 'Success (ground pad)';
- **First_Successful_Landing =** 2015-12-22
- SQL query identifying the date of the first successful ground landing.

# Successful Drone Ship Landing with Payload between 4000 and 6000

- %%sqlSELECT "Booster_Version"FROM SPACEXTABLEWHERE "Landing_Outcome" = 'Success (drone ship)'AND "PAYLOAD_MASS__KG_" > 4000AND "PAYLOAD_MASS__KG_" < 6000;
- SQL query listing booster versions that successfully landed on a drone ship with payloads between 4000 and 6000 kg.

| Booster_Version |
| ---: |
| F9 FT B1022 |
| F9 FT B1026 |
| F9 FT B1021.2 |
| F9 FT B1031.2 |

# Total Number of Successful and Failure Mission Outcomes

- %%sqlSELECT "Landing_Outcome", COUNT(*) AS Total_CountFROM SPACEXTABLEGROUP BY "Landing_Outcome“;
- SQL query showing a breakdown of landing outcomes, detailing successful and failed missions.

| Landing_Outcome | Total_Count |
|---:|---:|
| Controlled (ocean) | 5 |
| Failure | 3 |
| Failure (drone ship) | 5 |
| Failure (parachute) | 2 |
| No attempt | 21 |
| No attempt | 1 |
| Precluded (drone ship) | 1 |
| Success | 38 |
| Success (drone ship) | 14 |
| Success (ground pad) | 9 |
| Uncontrolled (ocean) | 2 |

# Boosters Carried Maximum Payload

- %%sqlSELECT "Booster_Version"FROM SPACEXTABLEWHERE "PAYLOAD_MASS__KG_" = (SELECT MAX("PAYLOAD_MASS__KG_") FROM SPACEXTABLE);
- SQL query listing boosters that carried the maximum payload.

| Booster_Version |
|---|
| F9 B5 B1048.4 |
| F9 B5 B1049.4 |
| F9 B5 B1051.3 |
| F9 B5 B1056.4 |
| F9 B5 B1048.5 |
| F9 B5 B1051.4 |
| F9 B5 B1049.5 |
| F9 B5 B1060.2 |
| F9 B5 B1058.3 |
| F9 B5 B1051.6 |
| F9 B5 B1060.3 |
| F9 B5 B1049.7 |
| |

# 2015 Launch Records

- %%sqlSELECT substr("Date", 6, 2) AS Month, "Booster_Version", "Launch_Site", "Landing_Outcome"FROM SPACEXTABLEWHERE "Landing_Outcome" = 'Failure (drone ship)'AND substr("Date", 0, 5) = '2015';
- SQL query showing launch outcomes of failed drone ship landings in 2015.

| Month | Booster_Version | Launch_Site | Landing_Outcome |
|---:|---:|---:|---:|
| 01 | F9 v1.1 B1012 | CCAFS LC-40 | Failure (drone ship) |
| 04 | F9 v1.1 B1015 | CCAFS LC-40 | Failure (drone ship) |

# Rank Landing Outcomes Between 2010-06-04 and 2017-03-20

- %%sqlSELECT "Landing_Outcome", COUNT(*) AS Outcome_CountFROM SPACEXTABLEWHERE "Date" BETWEEN '2010-06-04' AND '2017-03-20'GROUP BY "Landing_Outcome"ORDER BY Outcome_Count DESC;
- SQL query ranking landing outcomes between 2010 and 2017 based on the number of occurrences.

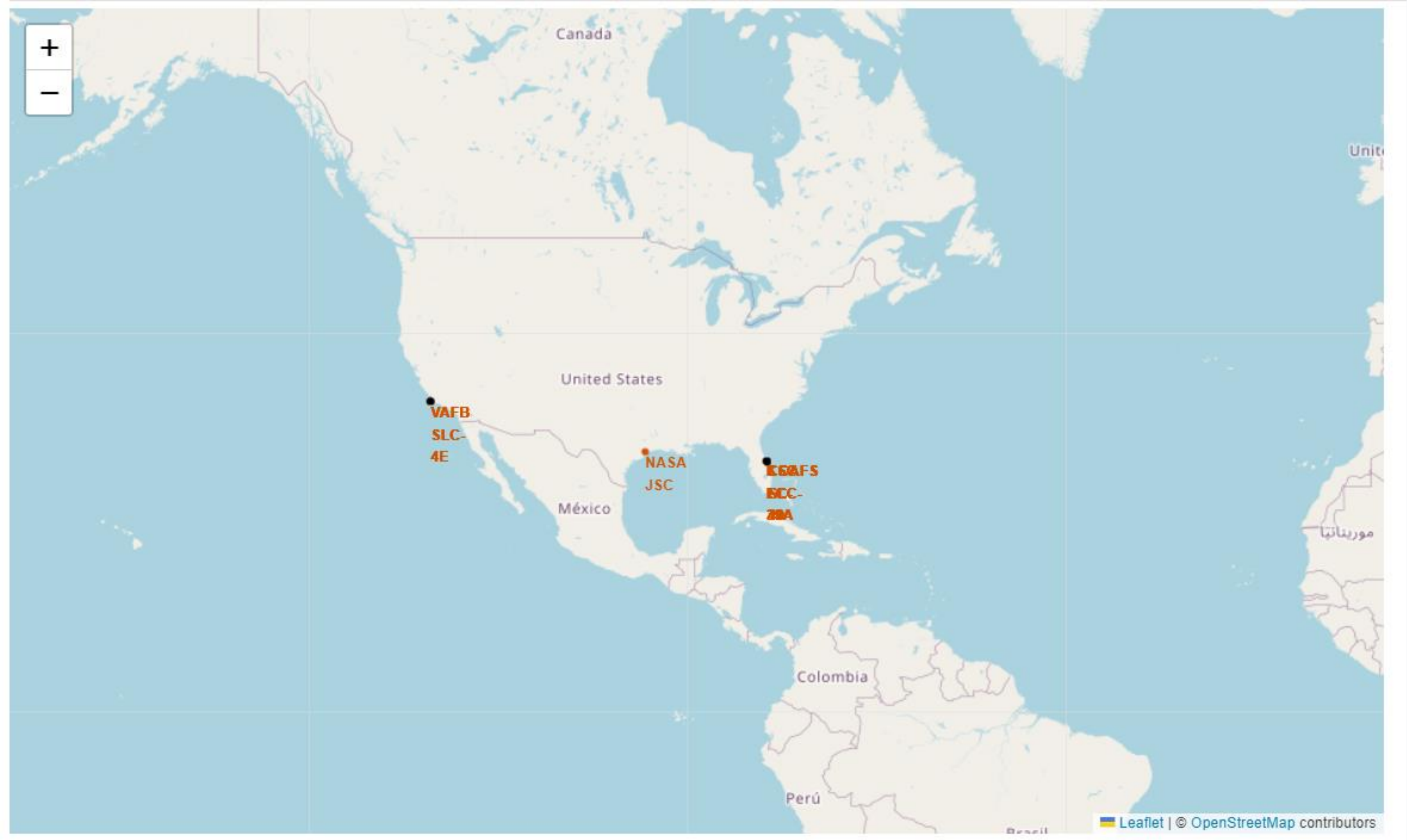
| Landing_Outcome | Outcome_Count |
|---:|---:|
| No attempt | 10 |
| Success (drone ship) | 5 |
| Failure (drone ship) | 5 |
| Success (ground pad) | 3 |
| Controlled (ocean) | 3 |
| Uncontrolled (ocean) | 2 |
| Failure (parachute) | 2 |
| Precluded (drone ship) | 1 |

Section 3

# Launch Sites Proximities Analysis

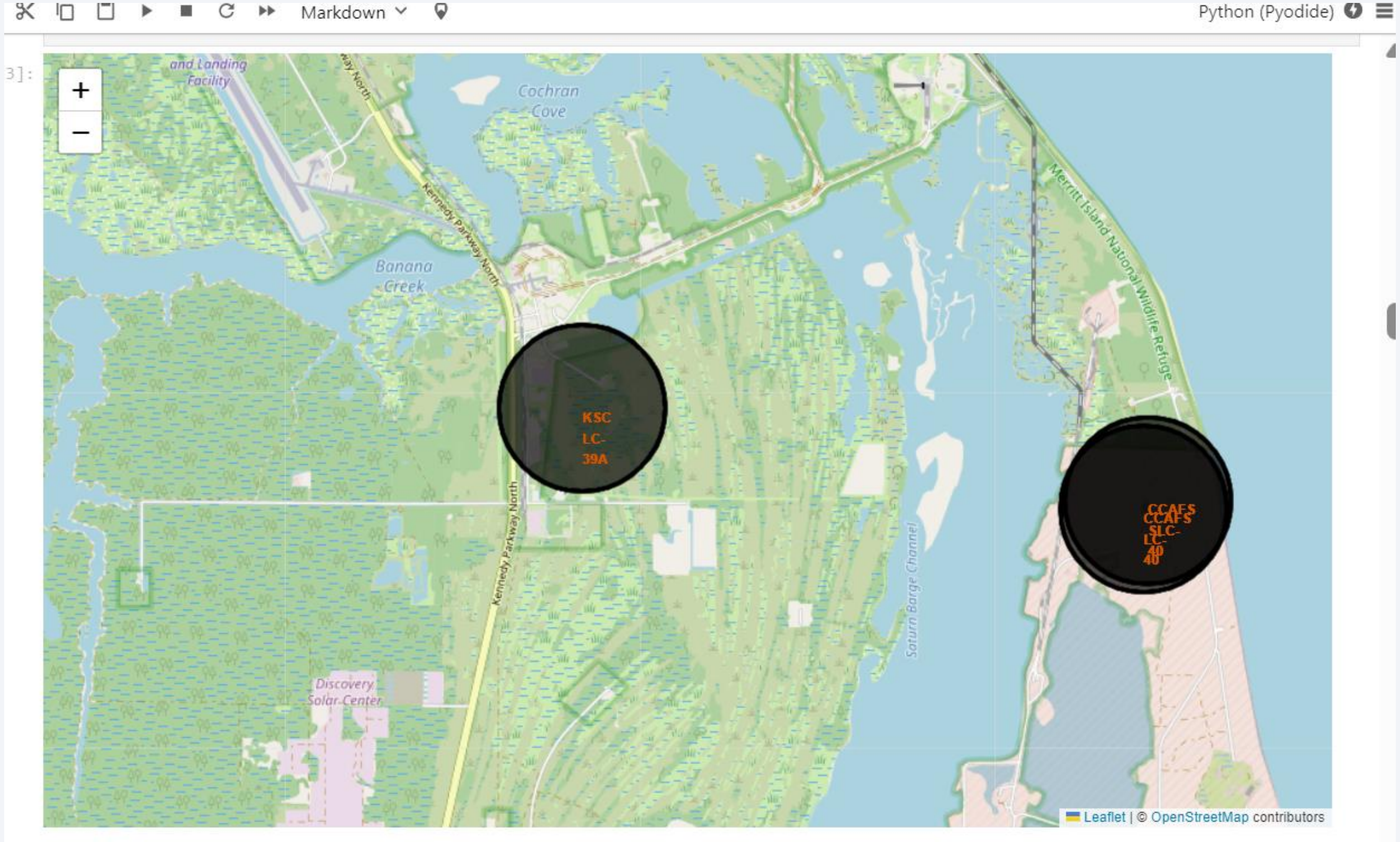
# <Folium Map Screenshot 1>

- Screenshot of a Folium map showing launch site locations and relevant markers.
- Proximity to equator is shown



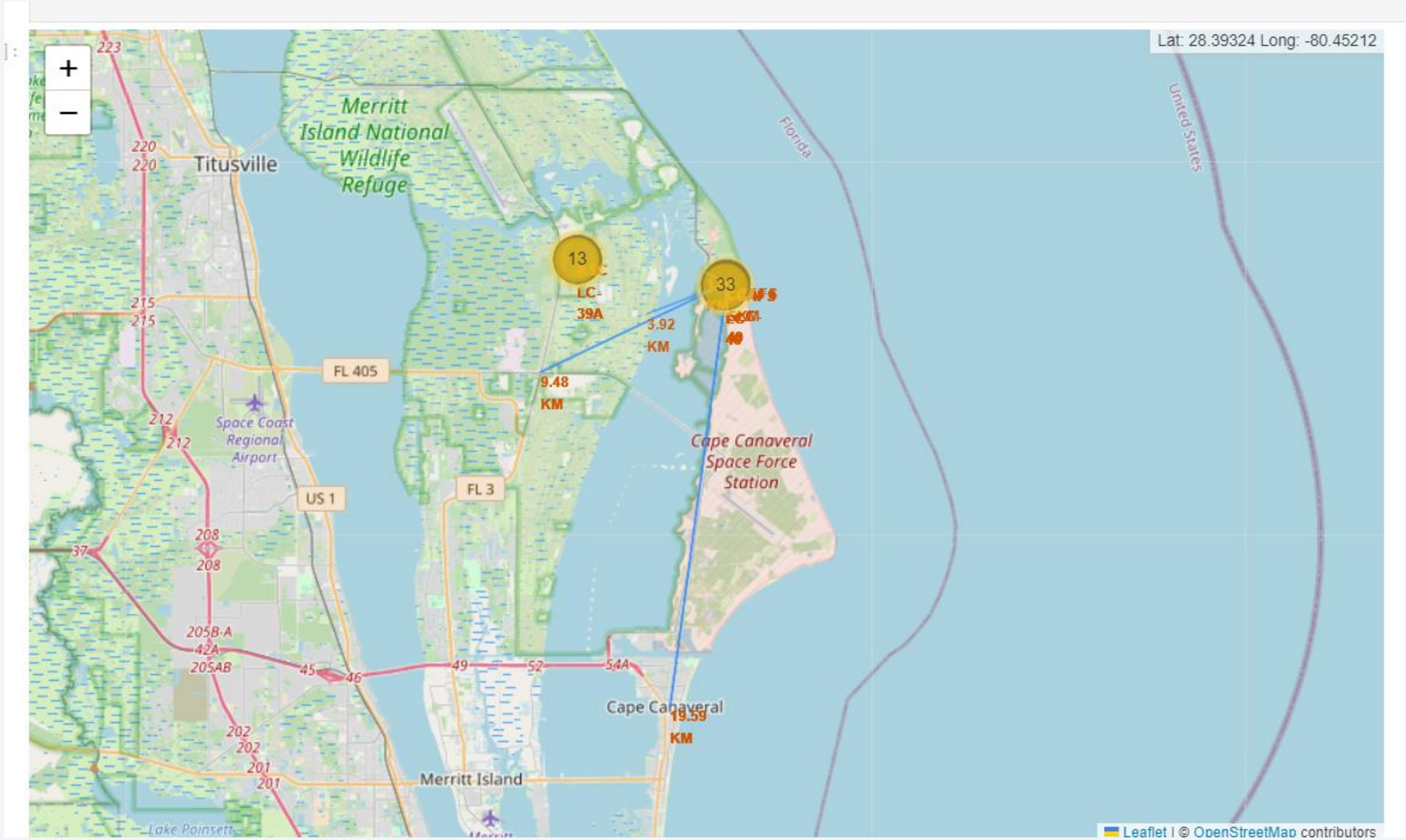
# <Folium Map Screenshot 2>

- Screenshot of a Folium map showing color-labeled launch outcomes.
- The cape caneveraval has highest launch activity



# <Folium Map Screenshot 3>

- Screenshot of a Folium map showing proximity of a selected launch site to infrastructure.
- Proximity to city , highway and railway is shown from CC

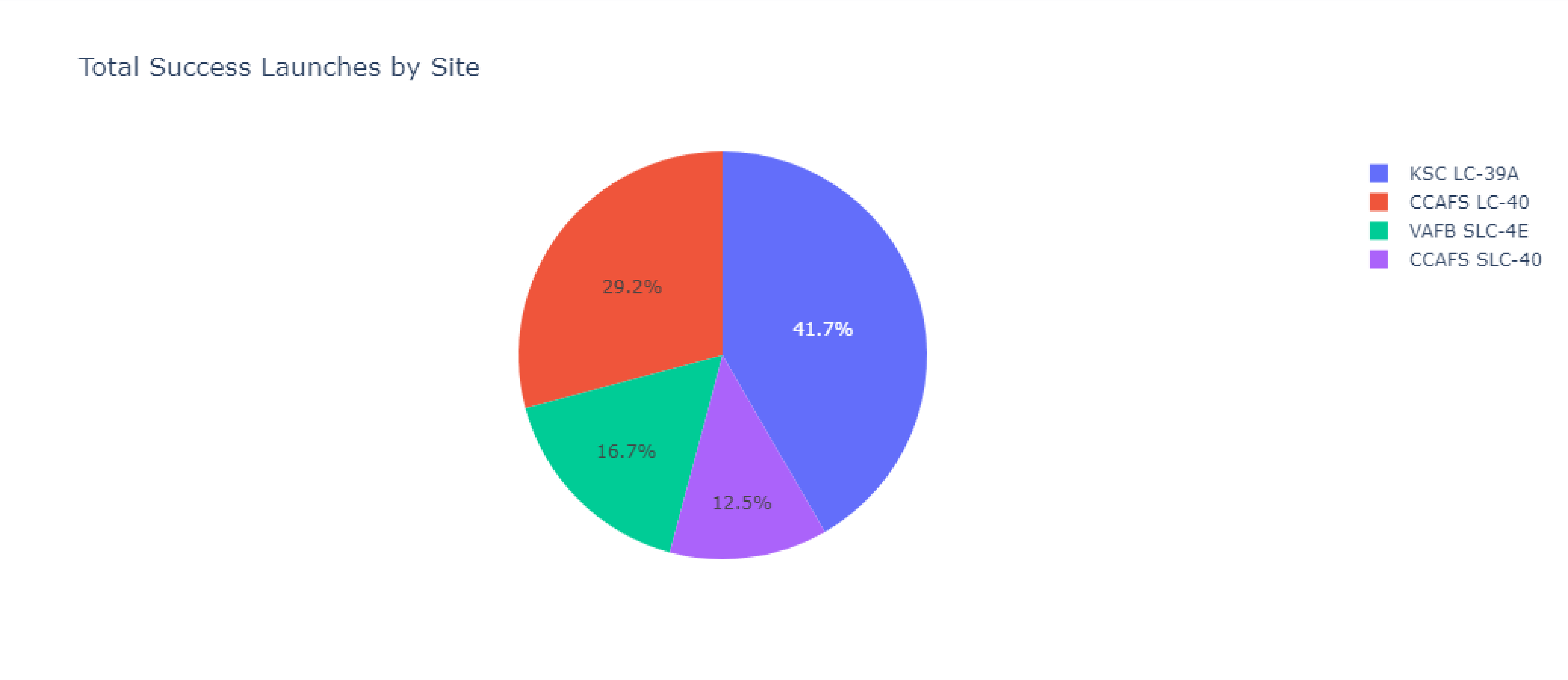


Section 4

# Build a Dashboard with Plotly Dash

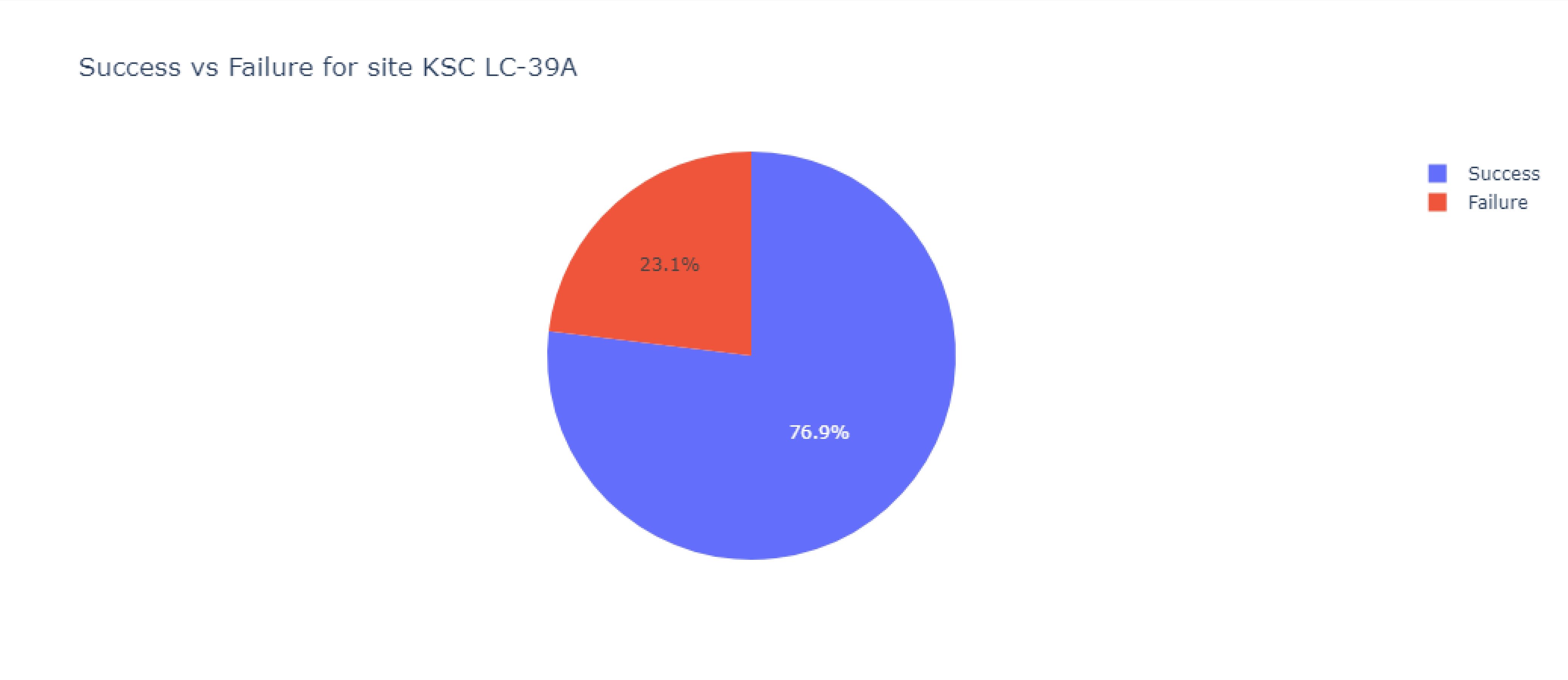
# <Dashboard Screenshot 1>

- Screenshot of a dashboard visualizing SpaceX launch success data interactively.



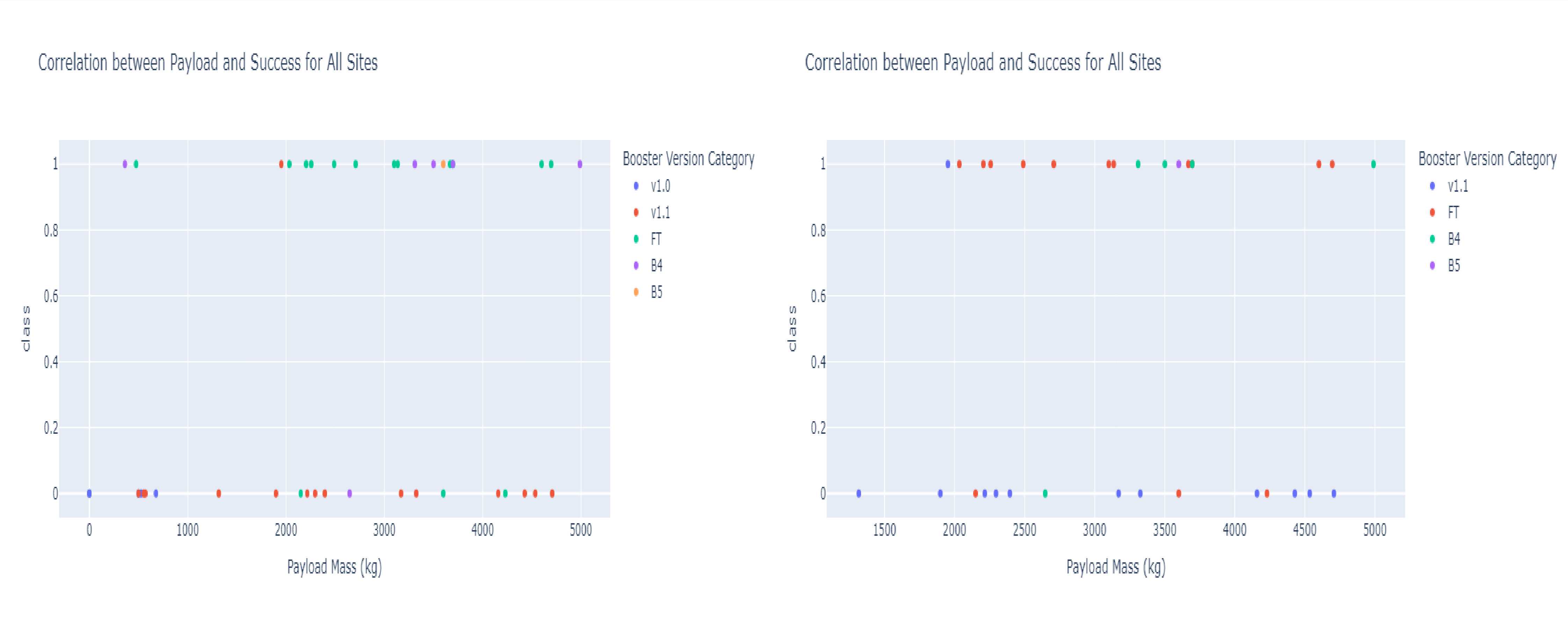
# <Dashboard Screenshot 2>

- Screenshot of a dashboard showcasing payload vs. success correlations.



# <Dashboard Screenshot 3>

- Screenshot of a dashboard showing launch site success rates interactively.

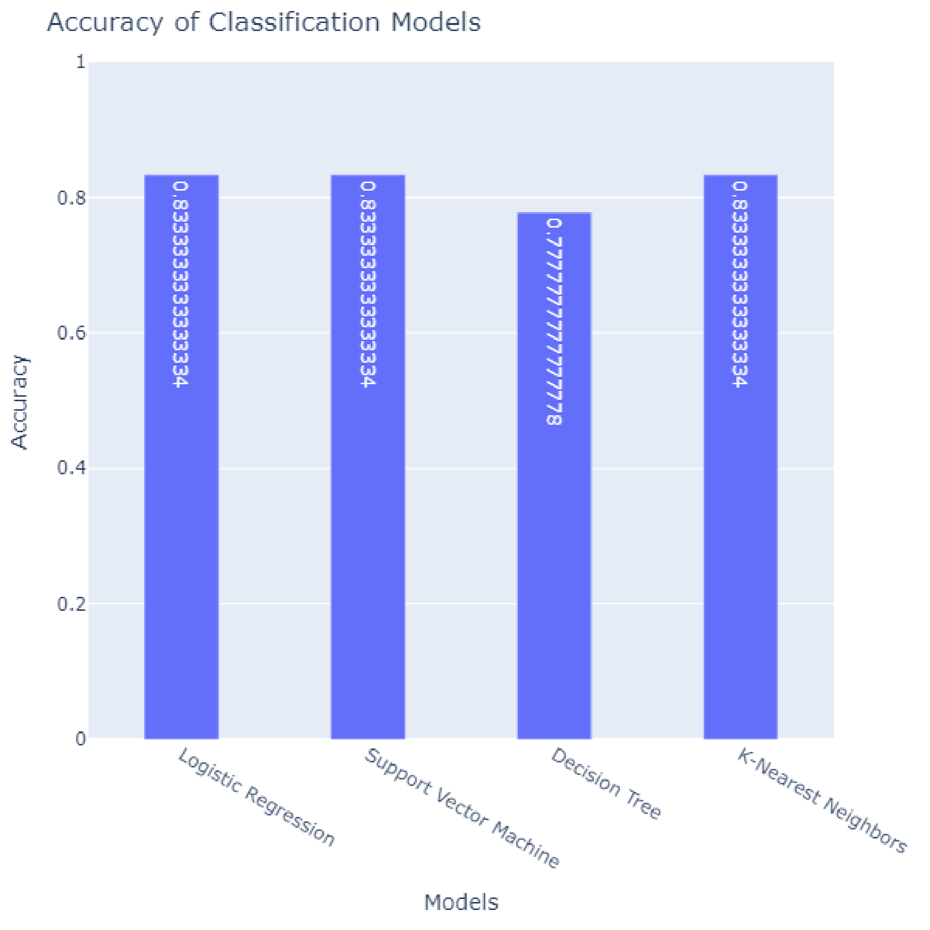


Section 5

# Predictive Analysis (Classification)

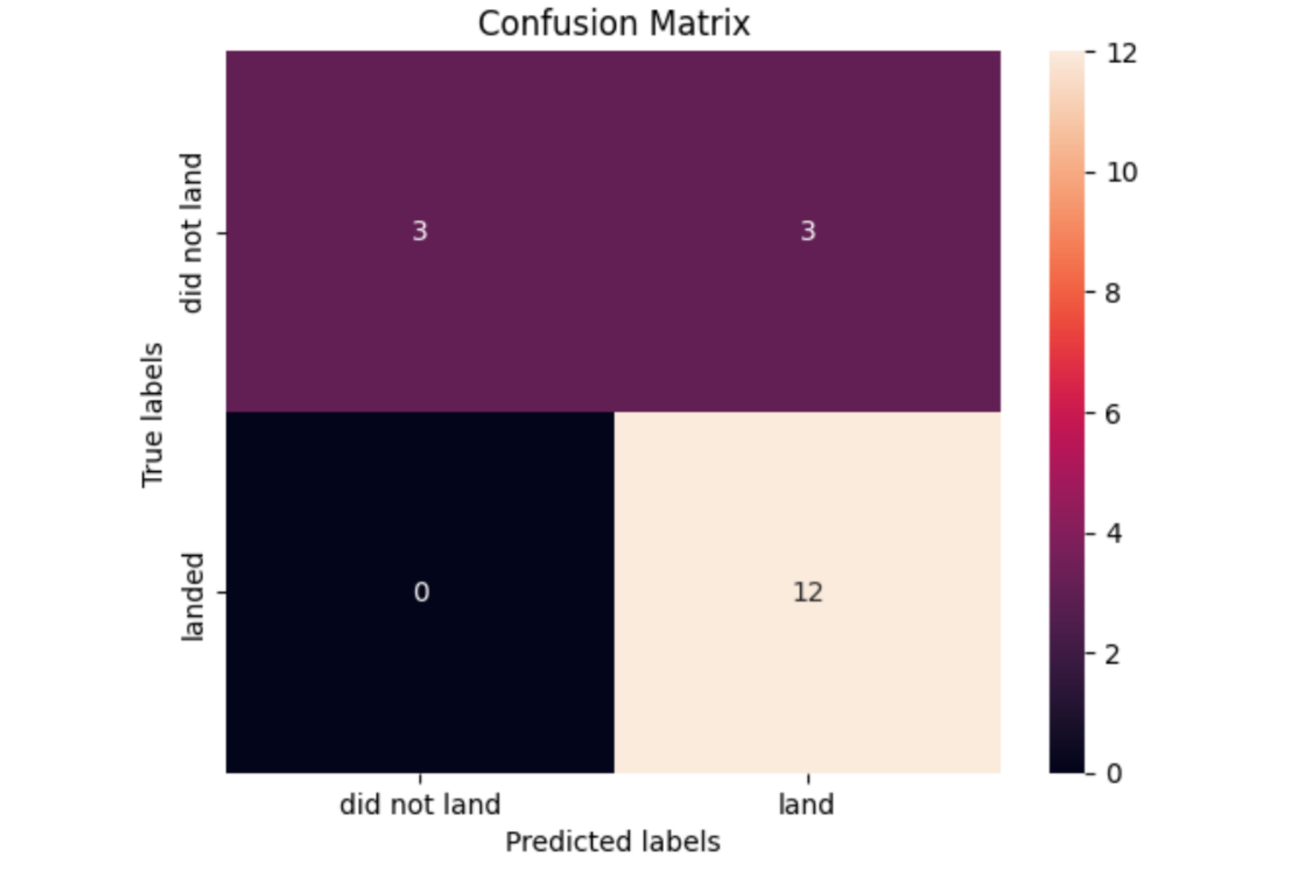
# Classification Accuracy

- Bar chart visualizing the accuracy of different classification models, showing SVM as the best performer.



# Confusion Matrix

- Confusion matrix displaying the performance of the SVM model, along with an explanation of its success in predicting outcomes.



# Conclusions

- **Insights**: Launch sites are strategically located near infrastructure (railways, highways) and coastlines, balancing operational efficiency with safety. Visualizations and interactive tools enhanced understanding of spatial and success factors.
- **Predictive Modeling**: SVM proved to be the most effective classifier, optimizing accuracy for predicting Falcon 9 first stage landings.
- **Next Steps**: Leverage these insights for operational improvements and refine models with additional data for better predictions.

# Appendix

- Git links
- SDSuhas/spacex-notebooks (github.com)

Thank you!